

CSC 2541: Machine Learning for Healthcare

Lecture 10: Generalization and Transfer Learning

Professor Marzyeh Ghassemi, PhD
University of Toronto, CS/Med
Vector Institute



Schedule

Jan 10, 2019, Lecture 1: Why is healthcare unique?

Jan 17, 2019, Lecture 2: Supervised Learning for Classification, Risk Scores and Survival

Jan 24, 2019, Lecture 3: Causal inference with observational data

Jan 31, 2019, Lecture 4: Fairness, Ethics, and Healthcare

Feb 7, 2019, Lecture 5: Clinical Time Series Modelling (Homework 1 due at 11:59 PM on MarkUs)

Feb 14, 2019, Lecture 6: Clinical Imaging (Project proposals due at 5PM on MarkUs)

Feb 21, 2019, Lecture 7: Clinical NLP and Audio

Feb 28, 2019, Lecture 8: Clinical Reinforcement Learning

Mar 7, 2019, Lecture 9: Missingness and Representations

Mar 14, 2019, Lecture 10: Generalization and transfer learning

Mar 21, 2019, Lecture 11: Interpretability / Humans-In-The-Loop / Policies and Politics

Mar 28, 2019, Course Presentations

April 4, 2019, Course Presentations

April 11, 2019, Project report due 11:59PM

Outline

1. Generalization
 - a. Why do we expect it?
 - b. Challenging in many settings!
2. Transfer Learning
 - a. Learning with other data?
 - b. Understanding what we move!
3. Why Do These Matter?
4. Project Discussion

Outline

1. **Generalization**

- a. Why do we expect it?
- b. Challenging in many settings!

2. Transfer Learning

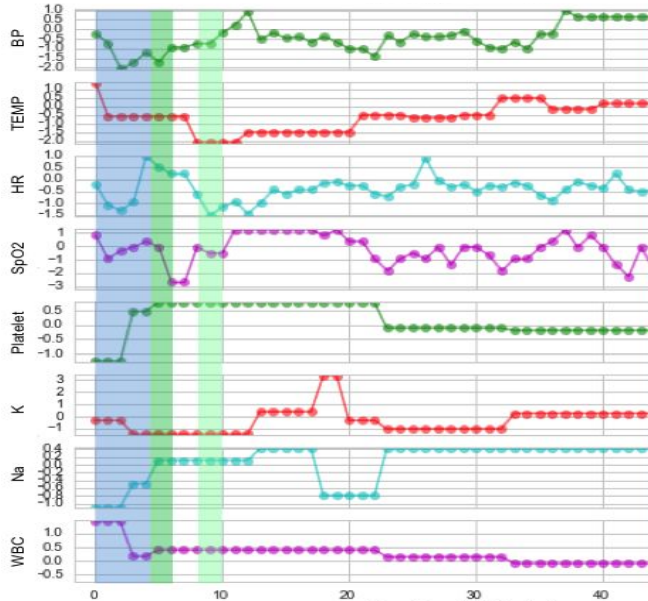
- a. Learning with other data?
- b. Understanding what we move!

3. Why Do These Matter?

4. Project Discussion

Remember? Hospital decision-making / care planning

Observe Patient Data

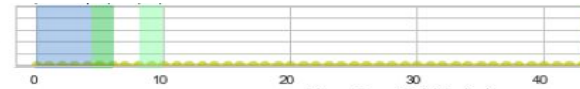


“Real-time” Prediction

Of {Drug/Mortality/Condition}

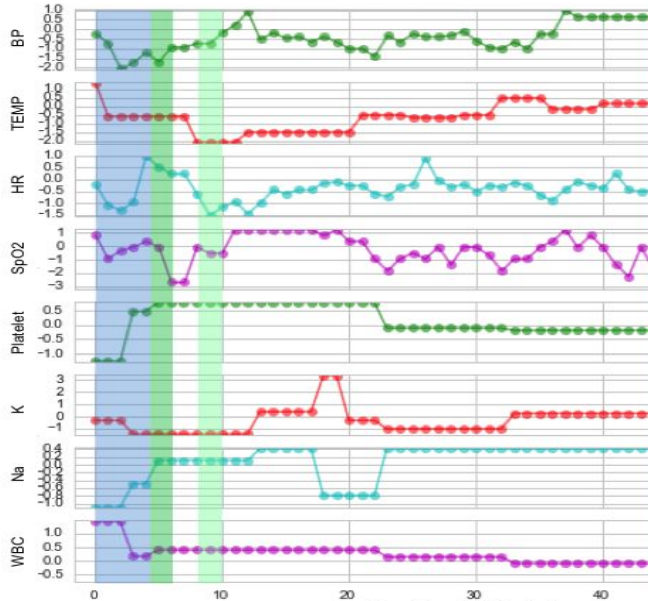
By Gap Time

?



Remember? Hospital decision-making / care planning

Observe Patient Data

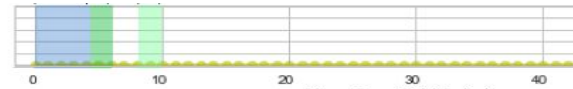


“Real-time” **Prediction**

Of {Drug/Mortality/Condition}

By Gap Time

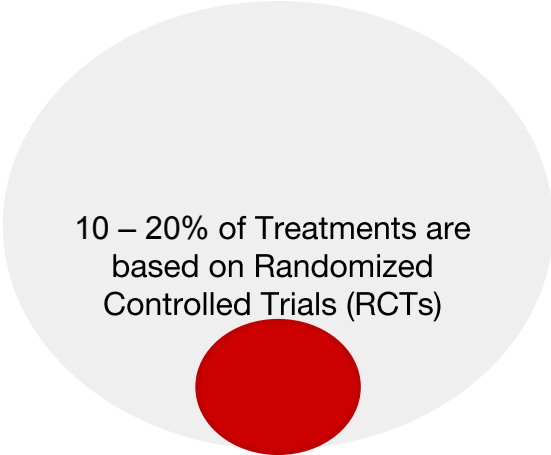
?



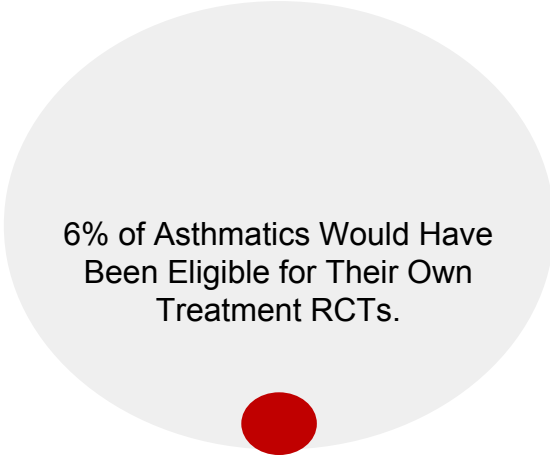
But Does This Generalize?

We Wanted More Generalizable Evidence in Health!

Randomized Controlled Trials (RCTs) are **rare and expensive**, and can encode **structural biases** that apply to very few people.



10 – 20% of Treatments are
based on Randomized
Controlled Trials (RCTs)



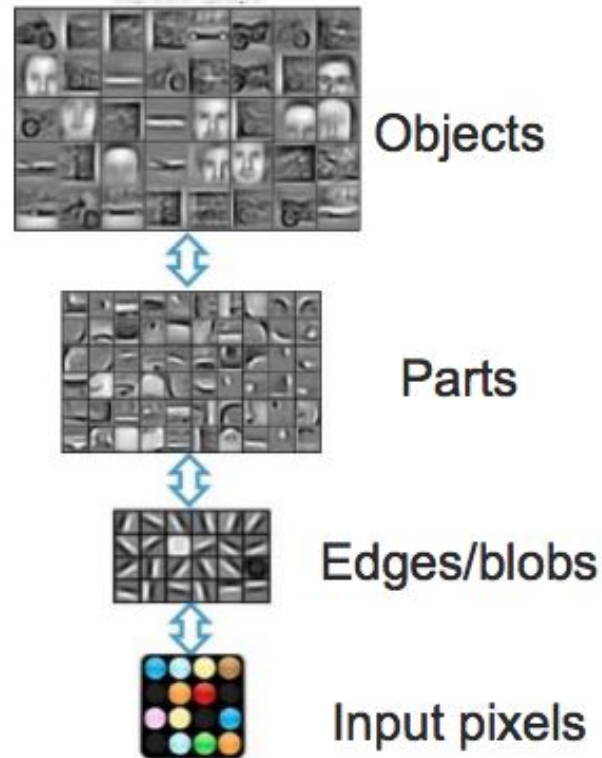
6% of Asthmatics Would Have
Been Eligible for Their Own
Treatment RCTs.

[1] Smith M, Saunders R, Stuckhardt L, McGinnis JM, Committee on the Learning Health Care System in America, Institute of Medicine. *Best Care At Lower Cost: The Path To Continuously Learning Health Care In America*. Washington: National Academies Press; 2013.

[2] Travers, Justin, et al. "External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?." *Thorax* 62.3 (2007): 219-223.

Why Do We Expect Generalization?

- We build predictive models because we **want** generalization.
- Intuition is that representation should be invariant to translation¹.
- Build models that learn generalizable hierarchies.



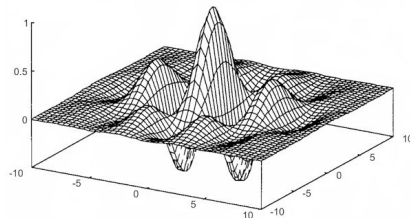
[1] L.P. Morency & T. Baltrusaitis, ACL 2017 Tutorial. <https://www.cs.cmu.edu/~morency/MMML-Tutorial-ACL2017.pdf>

Intuition for Image Generalization

- Objects can appear anywhere in a 2D image.
- Slide same “small window” with fixed weights across entire image.
- Each output value depends on small subset of input.
- Efficient learning - detect same pattern in any position in the image.
- Similar arguments for Bi-LSTMs in NLP.

Sidebar; What Are Those Lines?

- Gabor filters are frequency-based texture discriminants¹.
- Models trained on ImageNet from random initialization quickly learn Gabor filters², but **medical images** may not reduce to them...



[1] A.G. Ramakrishnan, S. Kumar Raja and H.V. Raghu Ram, "Neural network-based segmentation of textures using Gabor features," Proc. 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 365 - 374, 2002.

[2] Transfusion: Understanding Transfer Learning with Applications to Medical Imaging. M Raghu, C Zhang, J Kleinberg, S Bengio. arXiv preprint arXiv:1902.07208

What Do We Mean By Generalization?

- Simplest form of model generalization is measuring model performance on a held-out test set.
- Does performance on the test set mean that the model also performs well on a new test set?
- What if we use the same data source and cleaning¹?

[1] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do ImageNet Classifiers Generalize to ImageNet? arXiv:1806.00451 [cs, stat], June 2018. arXiv: 1806.00451.

What Are We Predicting?

- Our goal is to learn model f on some underlying true data distribution \mathcal{D}

$$L_{\mathcal{D}}(\hat{f}) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{I}[\hat{f}(x) \neq y]]$$

- Approximate it over a test set, S :

$$L_S(\hat{f}) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{I}[\hat{f}(x) \neq y]$$

- What could lead to loss differences between S and S' ?

$$L_S - L_{S'} = \underbrace{(L_S - L_{\mathcal{D}})}_{\substack{\text{Adaptivity gap} \\ \text{overfitting}}} + \underbrace{(L_{\mathcal{D}} - L_{\mathcal{D}'})}_{\substack{\text{Distribution Gap} \\ \text{biased selection}}} + \underbrace{(L_{\mathcal{D}'} - L_{S'})}_{\substack{\text{Generalization gap} \\ \text{random sampling error}}}$$

Do ImageNet Classifiers Generalize to ImageNet?

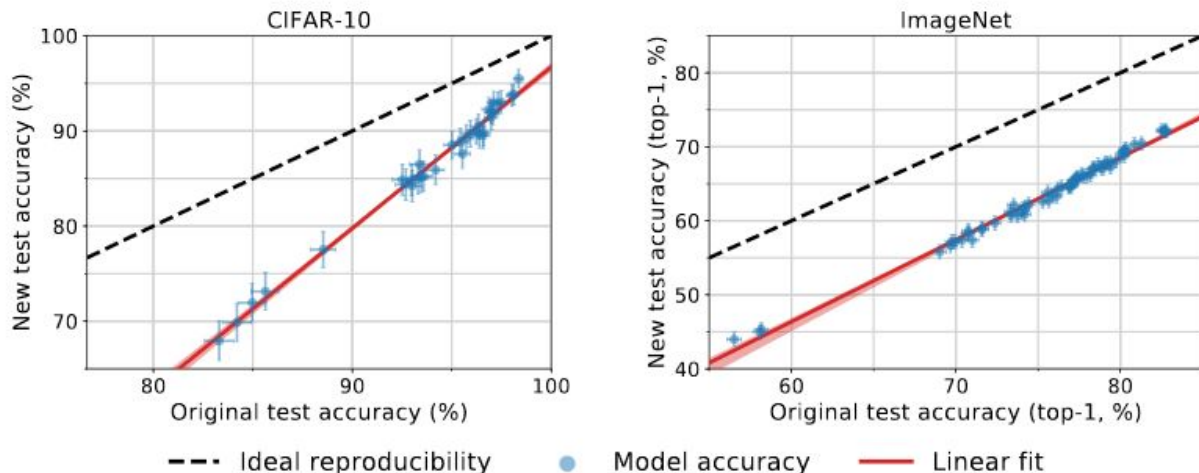


Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function with slope *greater* than 1 (1.7 for CIFAR-10 and 1.1 for ImageNet). This means that every percentage point of progress on the original test set translates into more than one percentage point on the new test set. The two plots are drawn so that their aspect ratio is the same, i.e., the slopes of the lines are visually comparable. The red shaded region is a 95% confidence region for the linear fit from 100,000 bootstrap samples.

Upshot?

“To put these accuracy numbers into perspective, we note that the best model in the ILSVRC 2013 competition achieved **89% top-5 accuracy**, and the best model from ILSVRC 2014 achieved **93% top-5 accuracy**.

So the 6% drop in top-5 accuracy from the 2018 state-of-the-art corresponds to **approximately five years of progress** in a very active period of machine learning research.”

Gap Sources

- Distributional gap is main source identified.

$$L_S - L_{S'} = \underbrace{(L_S - L_{\mathcal{D}})}_{\substack{\text{Adaptivity gap} \\ \text{"overfitting"}}} + \underbrace{(L_{\mathcal{D}} - L_{\mathcal{D}'})}_{\substack{\text{Distribution Gap} \\ \text{"biased selection"}}} + \underbrace{(L_{\mathcal{D}'} - L_{S'})}_{\substack{\text{Generalization gap} \\ \text{"random sampling error"}}$$

- Very hard to accurately replicate current image classification datasets distributions.
 - Human annotation is subjective.
 - Labels vary by annotator population, the exact task format, and compensation.
 - No exact definitions for many classes in ImageNet.

Gap Sources

- Distributional gap is main source identified.

$$L_S - L_{S'} = \underbrace{(L_S - L_{\mathcal{D}})}_{\substack{\text{Adaptivity gap} \\ \text{"overfitting"}}} + \underbrace{(L_{\mathcal{D}} - L_{\mathcal{D}'})}_{\substack{\text{Distribution Gap} \\ \text{"biased selection"}}} + \underbrace{(L_{\mathcal{D}'} - L_{S'})}_{\substack{\text{Generalization gap} \\ \text{"random sampling error"}}$$

- Very hard to accurately replicate current image classification datasets distributions.
 - **Human** annotation is **subjective**.
 - **Labels vary** by annotator population, the exact task format, and compensation.
 - **No exact definitions** for many classes in ImageNet.

Where Else Might This Be an Issue?

Media Advisory Wednesday, September 27, 2017

NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community

The dataset of scans is from more than 30,000 patients, including many with advanced lung disease.



What

The NIH Clinical Center recently released over 100,000 anonymized chest x-ray images and their corresponding data to the scientific

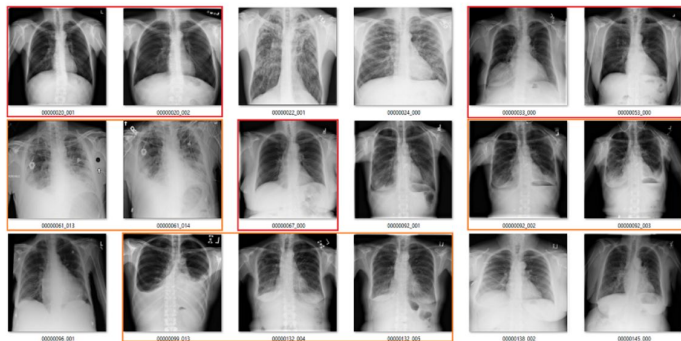


- 30,000+ patients and 100,000 images were released publicly by the NIH in 2017!
- Each image is given with 14 labels, and the goal is to predict them.

Similar Issues Hold

- Stanford ML Group creates ChexNet¹.
- Prominent radiologist manually reviews images and disagrees with labels².

Fibrosis



Our model, CheXNet, is a 121-layer convolutional neural network that inputs a chest X-ray image and outputs the probability of pneumonia along with a heatmap localizing the areas of the image most indicative of pneumonia.

We train CheXNet on the recently released ChestX-ray14 dataset, which contains 112,120 frontal-view chest X-ray images individually labeled with up to 14 different thoracic diseases, including pneumonia. We use dense connections and batch normalization to make the optimization of such a deep network tractable.



Input
Chest X-Ray Image

CheXNet
121-layer CNN

Output
Pneumonia Positive (85%)



We already saw above that the fibrosis labels are low accuracy, even being generous. But again, the problem is worse. In this image, the reds are incorrect labels, but the orange labels are where I have no idea. There are pleural effusions and/or consolidation. Could there be fibrosis under that? Sure, but there is no way to tell on these pictures.

[1] Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225 (2017).

[2] Oakden-Rayner, L. (2017). Exploring the chestxray14 dataset.

Slide credit to Irene Chen of MIT.

Mythos of Model Generalization?

- Training and test distributions often diverge for many reasons in practice.
- Detect with covariate shift methods, species estimation techniques, or sample selection bias learning.
- Often depends on ability to generate/use larger data.
- In medical setting, could this be a solution?

Other Sources?

CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilicus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, Andrew Y. Ng

(Submitted on 21 Jan 2019)

MIMIC-CXR: A large publicly available database of labeled chest radiographs

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, Steven Horng

(Submitted on 21 Jan 2019 (v1), last revised 23 Jan 2019 (this version, v2))

PadChest: A large chest x-ray image dataset with multi-label annotated reports

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, Maria de la Iglesia-Vayá

(Submitted on 22 Jan 2019 (v1), last revised 7 Feb 2019 (this version, v2))

- Each dataset is from a distinct institution.
- Different devices were used for collection.
- NIH clinical center gets “complicated” cases.
- Stanford/BIDMC are tertiary medical centers.
- No instances of emphysema in MIMIC-CXR.

Outline

1. Generalization
 - a. Why do we expect it?
 - b. Challenging in many settings!
2. **Transfer Learning**
 - a. Learning with other data?
 - b. Understanding what we move!
3. Why Do These Matter?
4. Project Discussion

Why Transfer Learning?

- Historically very few large sources of data.
- Transfer learning goal was to transfer model learning from large corpus to a smaller one¹.
- Also help to move into the correct areas of the loss landscape quickly with "winning tickets" subnetworks².



Supervised Classification



Semi-supervised Learning

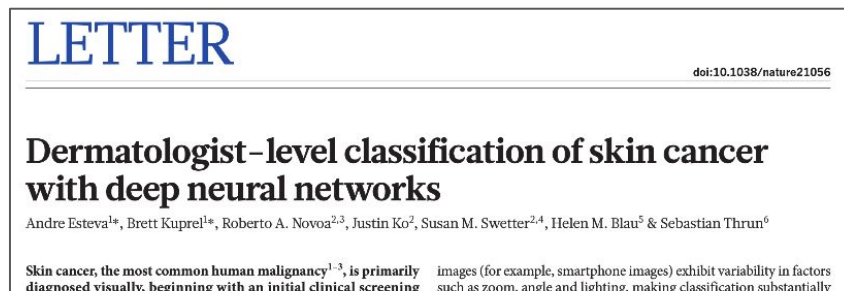


Transfer Learning

[1] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: Transfer learning from unlabeled data. ICML, 2007.

[2] Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." (2018).

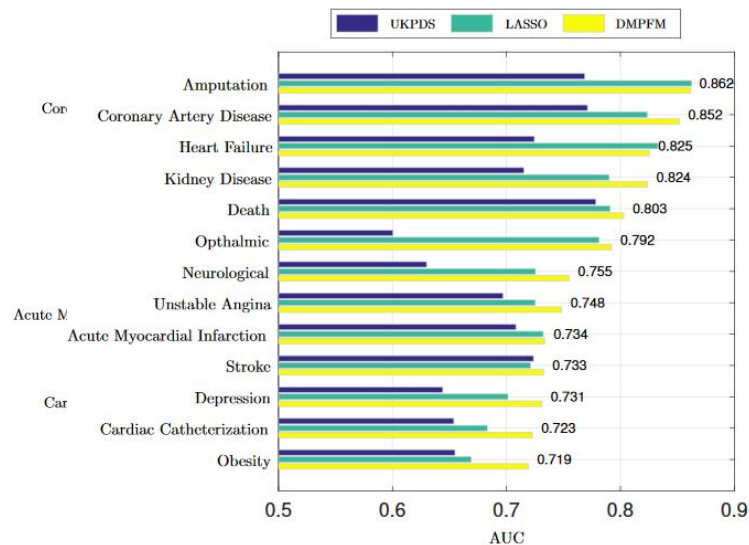
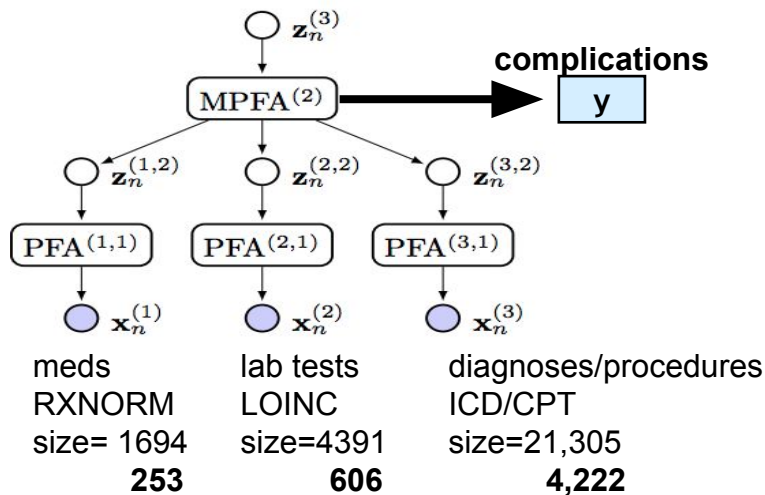
Operationally... ImageNet to Melanoma?



- Download ResNet18/50, VGG16/19, AlexNet, GoogLeNet, Inception-V3, ...
- Delete prior loss output layer, and replace with layer for melanoma prediction.
- Train on your (smaller) dataset - entire NN, last few layers, or the loss layer.

Could Also Learn Jointly

- Capture joint structure that leads to different EHR data with Deep Multi-task Poisson factor model + predict 13 complications, e.g., cardiovascular disease¹.

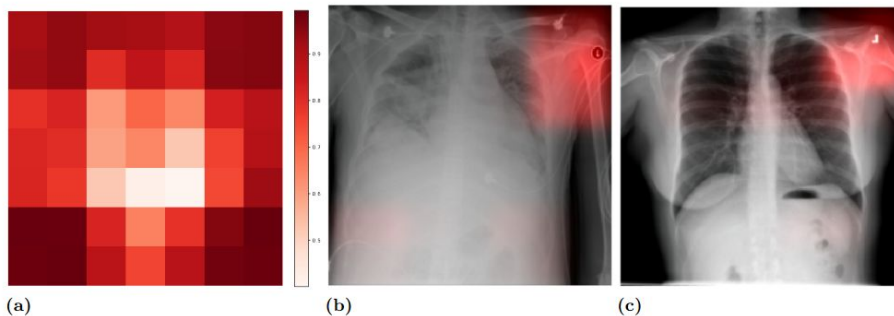


[1] Henao, Ricardo, et al. "Electronic health record analysis via deep poisson factor models." The Journal of Machine Learning Research 17.1 (2016): 6422-6453.

Problems Here Too

- CNN models can determine the hospital that the patient was admitted to from the X-ray¹.

Fig 4. CNN to predict hospital system detected both general and specific image features. (a) We obtained activation heatmaps from our trained model and averaged over a sample of images to reveal which subregions tended to contribute to a hospital system classification decision. Many different subregions strongly predicted the correct hospital system, with especially strong contributions from image corners. (b) On individual images, which have been normalized to highlight only the most influential regions and not all those that contributed to a positive classification, we note that the CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image field of view at the time they capture the image. When these strong features are correlated with disease prevalence, models can leverage them to indirectly predict disease.



[1] Zech, John R., et al. "Confounding variables can degrade generalization performance of radiological deep learning models." *arXiv preprint arXiv:1807.00431* (2018).

Issues With Transfer

- Encodes an explicit bias towards the solution learned on the source task¹.
- High-capacity models can learn to cheat using information you might not want embedded².



Figure 1: Details in x are reconstructed in GFx , despite not appearing in the intermediate map Fx .

[1] Li, Xuhong, Yves Grandvalet, and Franck Davoine. "Explicit inductive bias for transfer learning with convolutional networks." arXiv preprint arXiv:1802.01483 (2018).

[2] Chu, Casey, Andrey Zhmoginov, and Mark Sandler. "CycleGAN, a master of steganography." arXiv preprint arXiv:1712.02950 (2017).

Other Considerations

- One-shot Learning¹ or Model-Agnostic Meta Learning (MAML)²
Can we learn to classify female patients correctly on a few examples of female patients, even if our larger dataset contains no female patients?
- Zero-shot learning³
Can we classify Zika correctly, even if our larger dataset contains no prior instances of Zika?

[1] Fei-Fei, Li, Rob Fergus, and Pietro Perona. "One-shot learning of object categories." IEEE transactions on pattern analysis and machine intelligence 28.4 (2006): 594-611.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400, 2017.

[3] Palatucci, Mark, et al. "Zero-shot learning with semantic output codes." Advances in neural information processing systems. 2009.

Outline

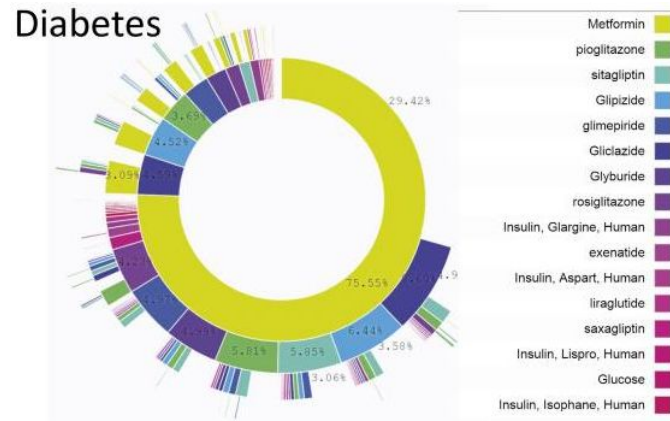
1. Generalization
 - a. Why do we expect it?
 - b. Challenging in many settings!
2. Transfer Learning
 - a. Learning with other data?
 - b. Understanding what we move!
3. **Why Do These Matter?**
4. Project Discussion

How Unique Do We Think People Are?

- Build an international data network with 11 data sources from four countries, including EHR and claims data on 250 million patients.
- Examine the treatment pathways for diabetes, depression and hypertension patients.
- How many followed a treatment pathway that was unique within the cohort, e.g., 0 nearest neighbors?

How Unique Do We Think People Are?

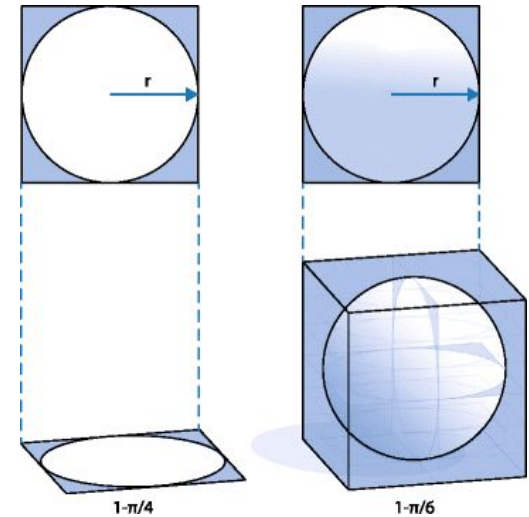
- 10% of diabetes patients, 11% of depression patients, and 24% of hypertension patients followed a treatment pathway that was **unique** within the cohort¹.



[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

What Does That Mean For Generalization and Transfer?

- High-dimensional data is fundamentally spread thinner.
- Ask “In an underlying population of 250 million, based on my 3-y treatment pathway, what patients are like me?”
For 24% of hypertension patients, “No one.”
- In high dimensions most of the volume of a sphere inscribed in a cube will be concentrated in its (many) corners.



Who Would Use These In Practice Anyways?

- US FDA has created Software Pre-certification (Pre-Cert) Pilot Program.
- Offers more flexibility and faster, iterative review processes.
- Establishes processes for software as medical device (SaMD) technologies, which may include software functions that use artificial intelligence and machine learning algorithms.

See: <https://hackernoon.com/demystifying-the-current-upward-trend-in-fda-approvals-of-medical-devices-using-artificial-cb9cc18d175>

FDA-Approved AI in 2018, Case 1

- The OsteoDetect software is a computer-aided detection and diagnostic software that uses “AI” to analyze two-dimensional X-ray images for signs of distal radius fracture.
- The company submitted a retrospective study of 1,000 radiograph images that assessed independent algorithm performance for detecting wrist fractures, and the accuracy of the fracture localization of OsteoDetect, against the performance of three board certified orthopedic hand surgeons.
- Demonstrated that the readers’ performance in detecting wrist fractures was improved using the software, Sens/Spec/PPVNPV, when aided by OsteoDetect, as compared with performance in standard clinical practice.

FDA-Approved AI in 2018, Case 2

- The IDX software is designed to detect greater than a mild level of diabetic retinopathy, which causes vision loss and affects 30 million people in the US. It occurs when high blood sugar damages blood vessels in the retina
- The program uses “AI” to analyze images of the adult eye taken with a special retinal camera. A doctor uploads the images to a cloud server, and the software then delivers a positive or negative result.
- The FDA based its decision on data from a clinical study of 900 diabetes patients’ retinal images collected from 10 primary care sites. IDx-DR correctly identified 87.4% of more severe diabetic retinopathy; images with mild/lesser diabetic retinopathy were correctly identified 89.5% of the time.

FDA-Approved AI in 2018, Case 3

- Viz's first product automatically analyzing CT scans of ER patients, and uses an "AI" to detect blockages in major brain blood vessels
- The company submitted a retrospective study of 300 CT images that assessed the independent performance of the image analysis algorithm and notification functionality of the Viz.AI Contact application against the performance of two trained neuro-radiologists for the detection of large vessel blockages in the brain.
- Real-world evidence was used with a clinical study to demonstrate that the application could notify a neurovascular specialist sooner in cases where a blockage was suspected.

FDA-Approved AI in 2018, Case 4

- Arterys is an “AI” cloud medical imaging software that helps clinicians measure and track tumors or potential cancers, and easily apply radiological standards.
- Automates the segmentation of lung nodules and liver lesions, with accuracy equal to segmentations performed manually by experienced clinicians. The clinician has the capability to edit these automated segmentations/remain in control.
- 510(K) approval through comparison of software identification and measurement of lesions to expert-assessed images and showing excellent correlation in liver MRI/CT scans, and lung CT scans.

Outline

1. Generalization
 - a. Why do we expect it?
 - b. Challenging in many settings!
2. Transfer Learning
 - a. Learning with other data?
 - b. Understanding what we move!
3. Why Do These Matter?
4. **Project Discussion**